

Bulk RNA-seq 差异表达分析报告

GSE147507 · NHBE · SARS-CoV-2 感染 vs 对照（公开数据示例）

项目 DEMO-RNA-01

数据来源 GEO GSE147507

分析日期 2026-10-01

出具 HST GENOMICS 生物信息分
析

项目概况

本报告是公开数据示例交付物，用于展示 HST GENOMICS bulk RNA-seq 差异表达分析的交付形式。数据来自 GEO 公开数据集 GSE147507 (Blanco-Melo 等, 2020)，选取其中 Series 1: 人原代支气管上皮细胞 (NHBE) SARS-CoV-2 感染组与对照组各 3 个样本的原始计数矩阵。

6	14,159	111	95 / 16
样本	过滤后保留的基因	差异基因 ($\text{padj} < 0.05$, $ \log_2\text{FC} > 1$)	上调 / 下调

样本	分组	计数总和
Series1_NHBE_Mock_1	对照	11.3 M
Series1_NHBE_Mock_2	对照	10.3 M
Series1_NHBE_Mock_3	对照	16.6 M
Series1_NHBE_SARS-CoV-2_1	SARS-CoV-2 感染	10.2 M
Series1_NHBE_SARS-CoV-2_2	SARS-CoV-2 感染	9.9 M
Series1_NHBE_SARS-CoV-2_3	SARS-CoV-2 感染	29.3 M

表 1 样本与分组

分析方法

- 输入：GEO 提供的基因水平原始计数矩阵（作者以 STAR 比对、featureCounts 计数），本示例不重新比对。
- 过滤：6 个样本计数总和 < 10 的基因剔除，保留 14,159 / 21,797 个基因；其中 13,061 个通过 DESeq2 独立过滤，参与多重检验校正。
- 差异表达：PyDESeq2 0.5.4（DESeq2 方法的 Python 实现），设计 ~condition，对比 infected vs mock，Wald 检验，Benjamini-Hochberg 校正；显著性阈值 $p_{adj} < 0.05$ 且 $|\log_2FC| > 1$ 。
- 质控：文库大小、基于前 500 个高变基因的主成分分析、样本间 Spearman 相关。
- 富集：上调基因做 MSigDB Hallmark 过代表性分析（Enrichr, gseapy 1.3.1）。

质量控制

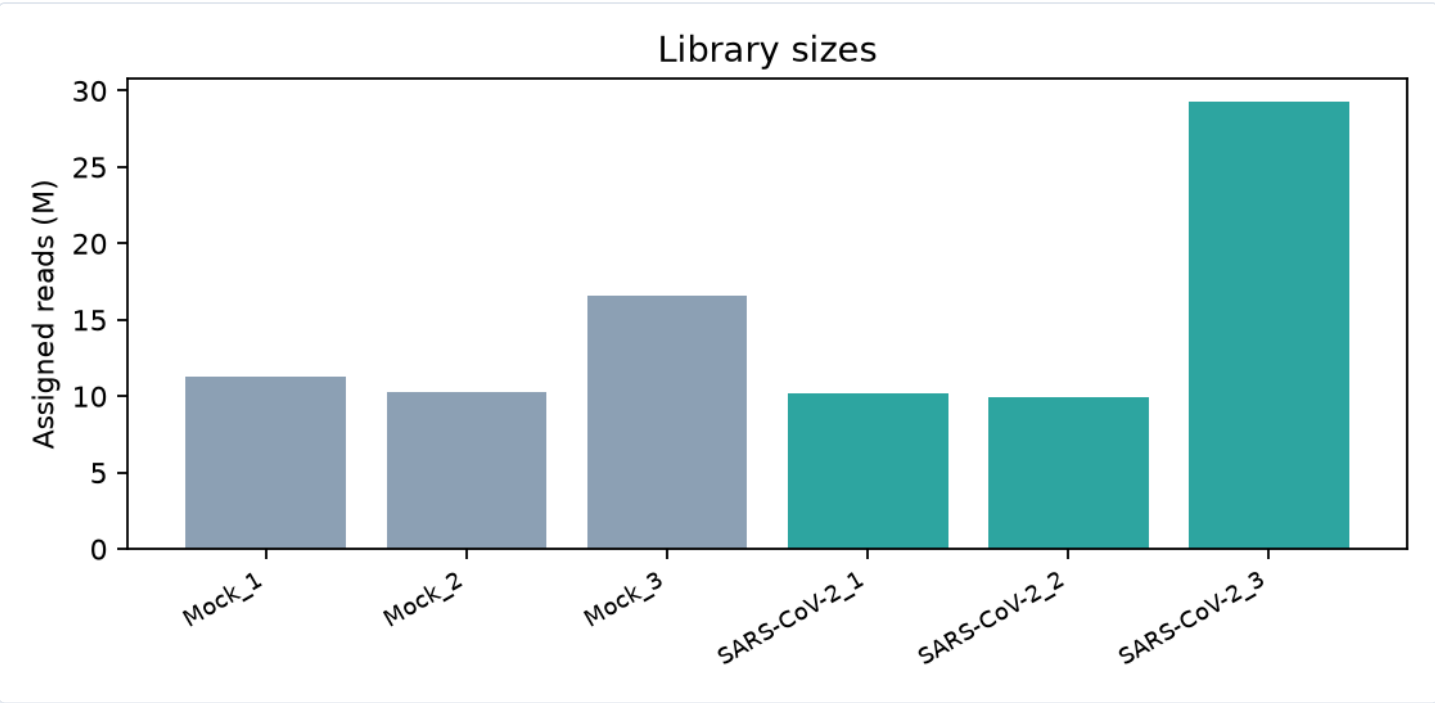


图 1 各样本计数总和。

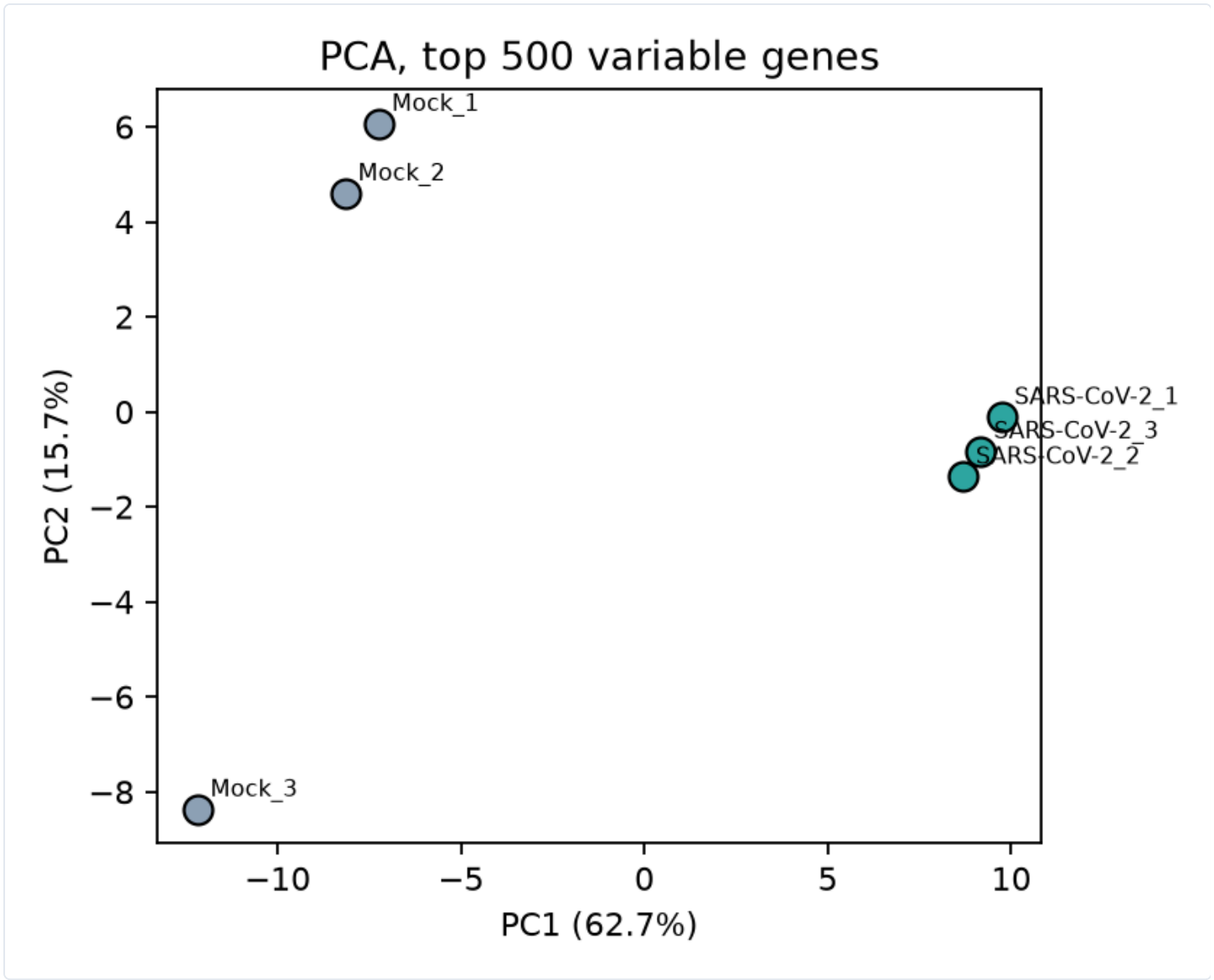


图 2 主成分分析。PC1 解释 62.7% 方差，感染组与对照组沿 PC1 分开。

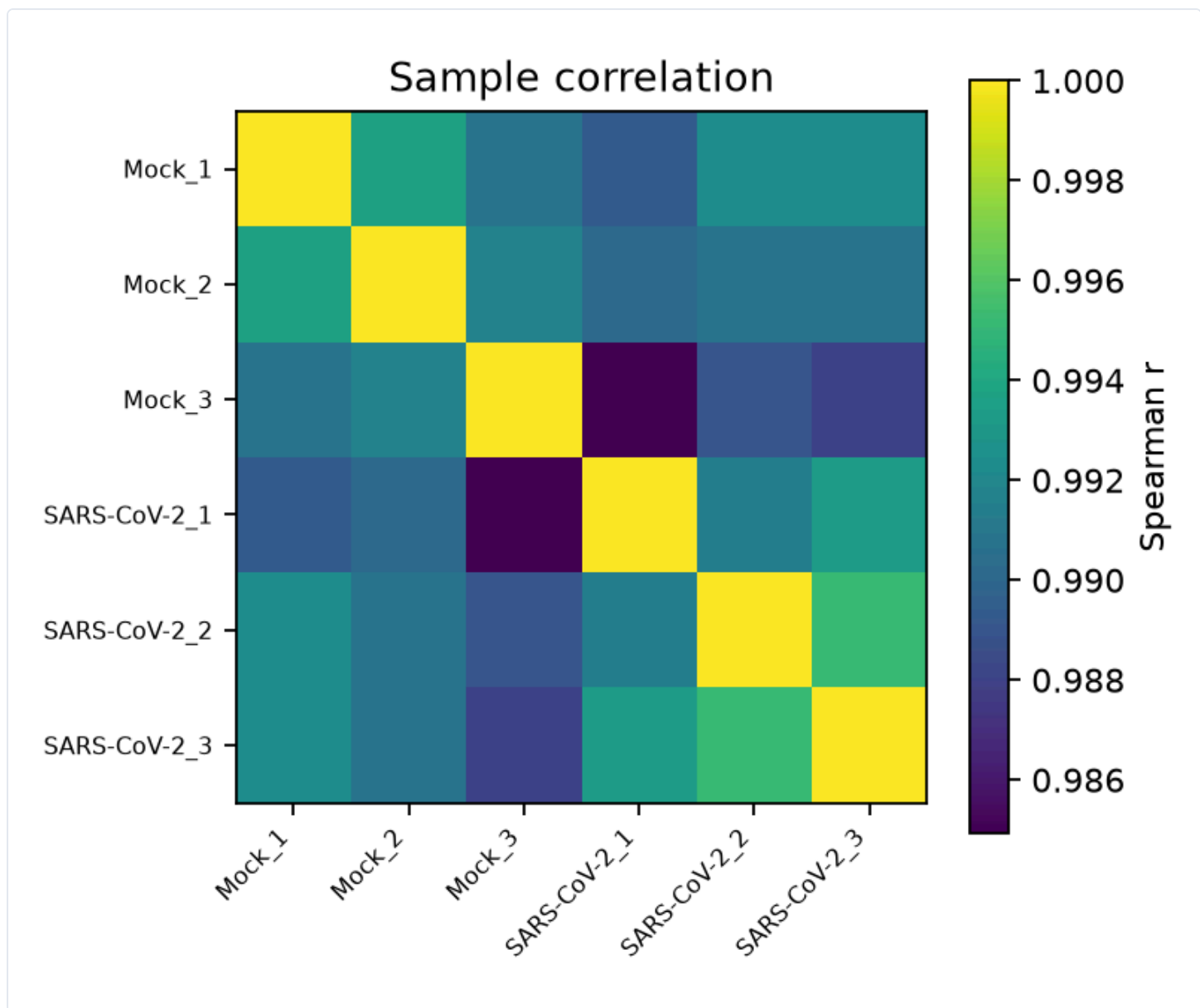


图 3 样本间 Spearman 相关系数。

结果

在 $\text{padj} < 0.05$ 且 $|\log_2\text{FC}| > 1$ 的阈值下共得到 111 个差异基因，其中上调 95 个、下调 16 个。

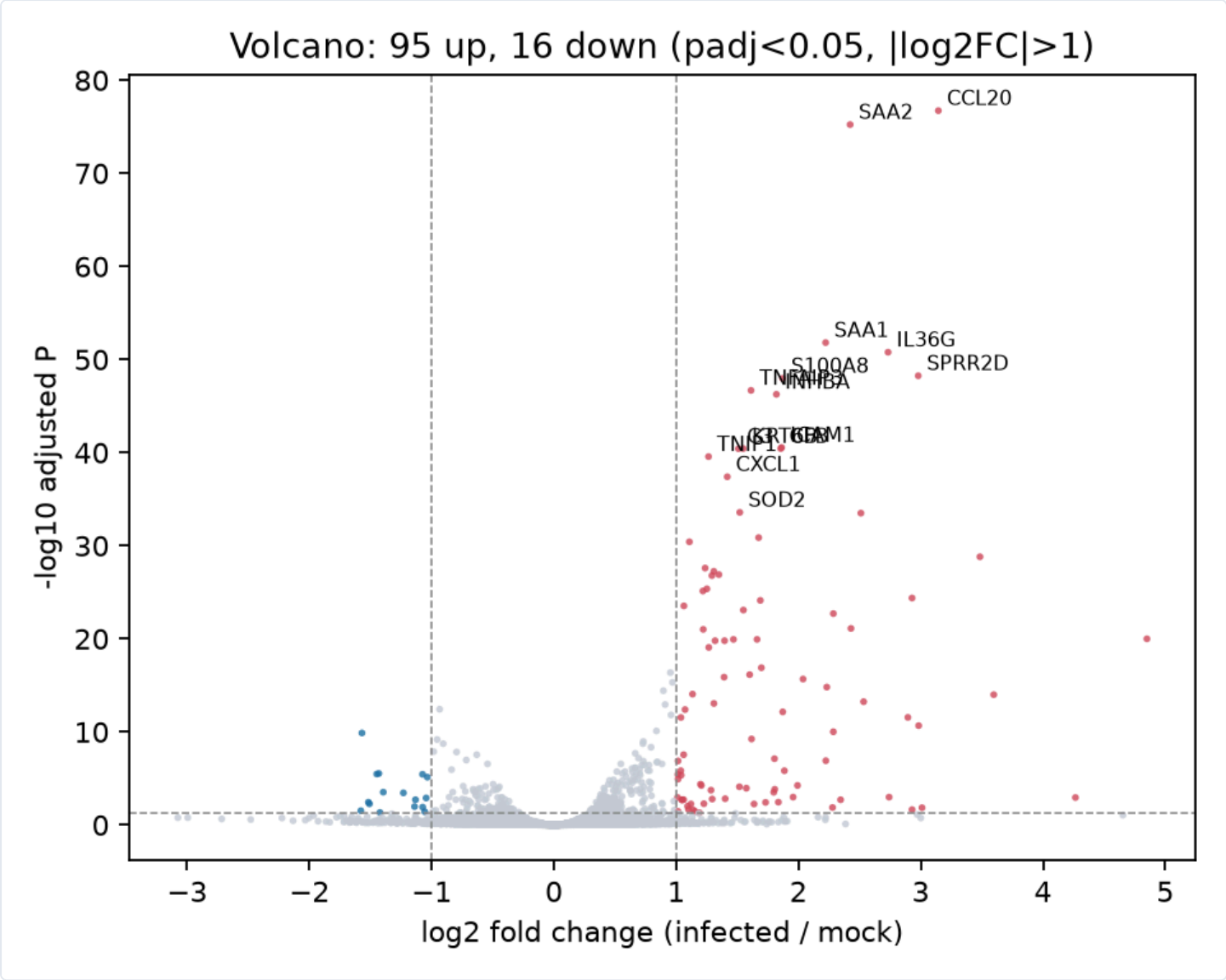


图 4 火山图，标注 padj 最小的 15 个基因。

基因	baseMean	log2FC	padj
CCL20	412.1	3.14	1.87e-77
SAA2	575.8	2.42	5.82e-76
SAA1	3316.1	2.22	1.54e-52
IL36G	271.3	2.73	1.65e-51
SPRR2D	365.2	2.98	5.60e-49
S100A8	1707.1	1.87	1.07e-48
TNFAIP3	2288.8	1.61	2.09e-47
INHBA	1216.9	1.82	5.53e-47
ICAM1	1884.7	1.86	2.81e-41
KRT6B	2803.4	1.55	3.71e-41

CFB	789.3	1.85	3.71e-41
C3	6336.0	1.51	3.85e-41
TNIP1	4659.3	1.26	2.71e-40
CXCL1	1801.7	1.42	3.96e-38
SOD2	2001.4	1.52	2.65e-34
MX1	427.7	2.51	3.19e-34
ZC3H12A	870.7	1.67	1.37e-31
S100A9	12313.2	1.11	3.89e-31
CXCL5	104.7	3.48	1.56e-29
C15orf48	897.0	1.24	2.54e-28

表 2 padj 最小的 20 个差异基因（完整表见 tables/DE_results_all_genes.csv）。

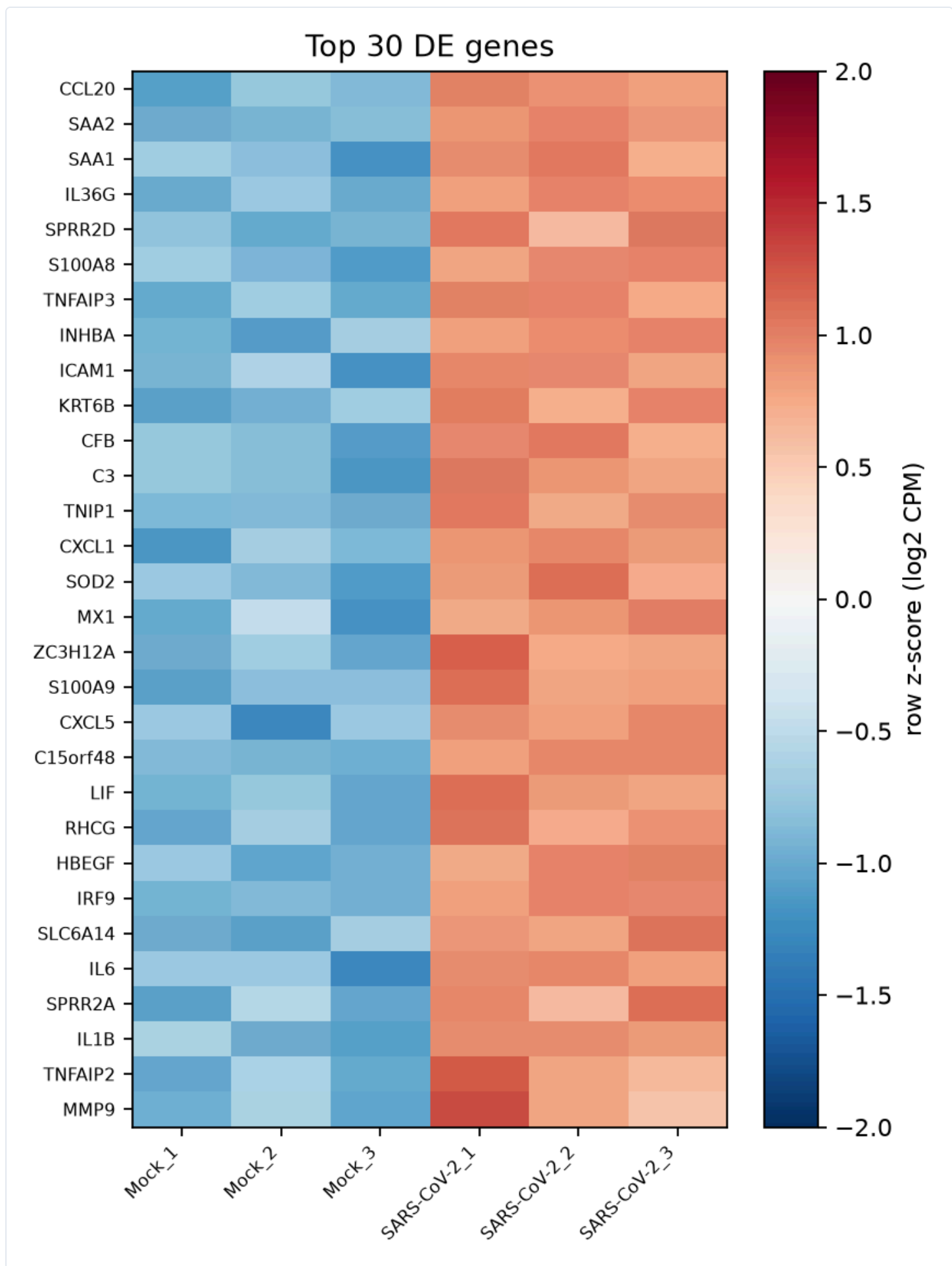


图 5 前 30 个差异基因的表达热图（行 z 分数）。

与原文对照：上调前 20 个基因中只有 1 个是经典干扰素刺激基因（MX1），列表以趋化因子与炎症基因为主（CCL20、CXCL1、CXCL5、IL36G 等），同时 Hallmark 干扰素 γ 与干扰素 α 应答通路在上调基因中显著富集（见表 3）：干扰素刺激基因有诱导，但幅度不如趋化因子突出。这与原文（Blanco-Melo 等，2020）报道的干扰素应答相对偏弱、趋化因子强烈诱导相符。

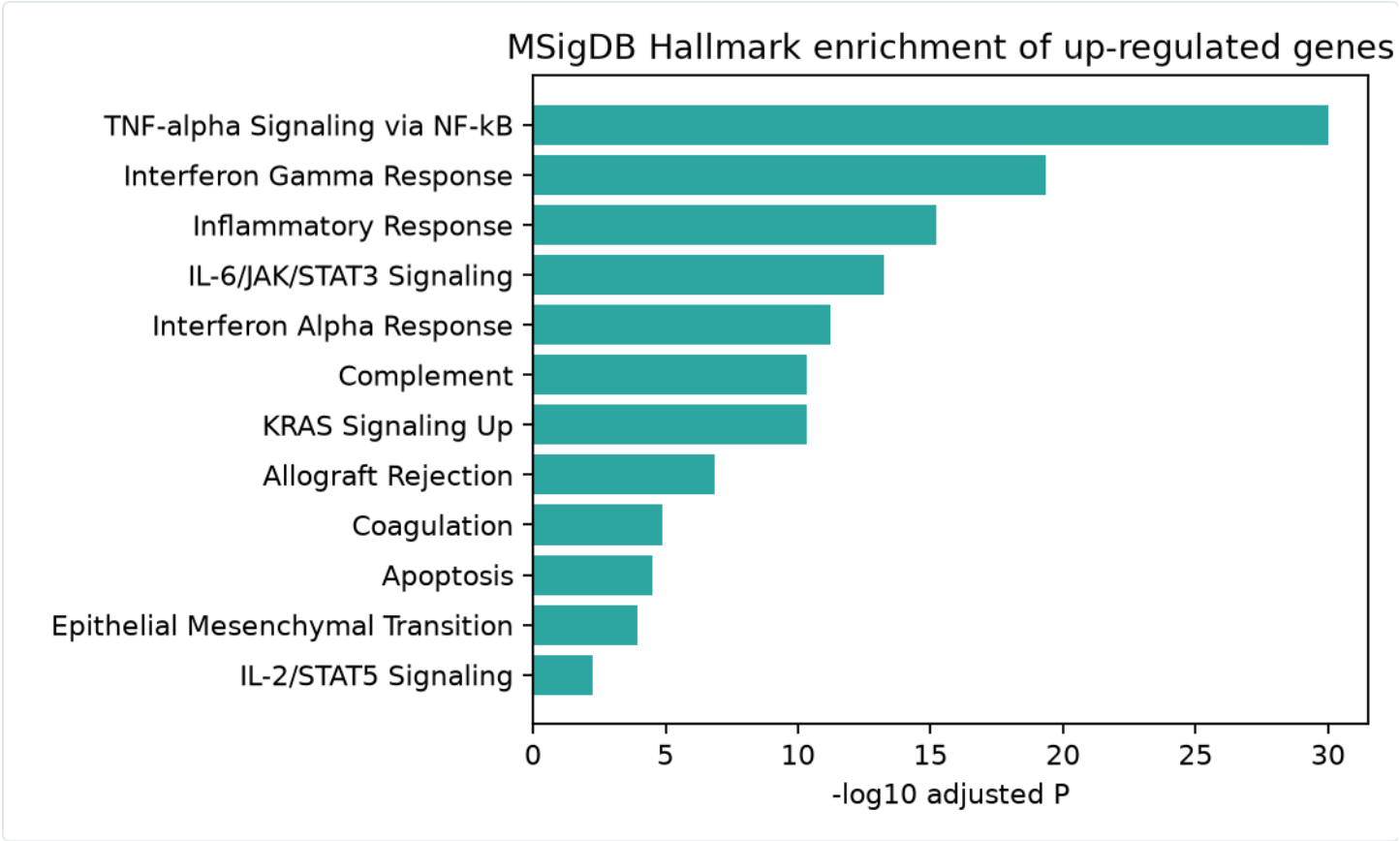


图 6 上调基因的 MSigDB Hallmark 富集。

通路	重叠	校正 P
TNF-alpha Signaling via NF-kB	27/200	9.77e-31
Interferon Gamma Response	20/200	4.74e-20
Inflammatory Response	17/200	6.07e-16
IL-6/JAK/STAT3 Signaling	12/87	5.97e-14
Interferon Alpha Response	11/97	6.10e-12
Complement	13/200	4.60e-11
KRAS Signaling Up	13/200	4.60e-11
Allograft Rejection	10/200	1.35e-07
Coagulation	7/138	1.33e-05
Apoptosis	7/161	3.28e-05
Epithelial Mesenchymal Transition	7/200	1.20e-04

表 3 富集结果前 12 条。

交付物清单

- tables/DE_results_all_genes.csv：全部基因的 DESeq2 结果
- tables/DE_significant_padj0.05_log2FC1.csv：显著差异基因
- tables/enrichment_hallmark_upregulated.csv：富集结果
- tables/counts_raw_selected_samples.csv、tables/sample_sheet.csv：输入计数与样本表
- figures/*.png：出版级图（可按期刊要求改为 PDF/SVG）
- run_bulk.py、report_lib.py、requirements.txt、README.txt：可复现的完整分析代码、参数与运行说明

方法学段落（可直接改写进论文）

基因水平原始计数来自 GEO（GSE147507）。剔除六个样本计数总和低于 10 的基因后，使用 PyDESeq2 0.5.4 进行差异表达分析。

本示例使用公开数据，所有数字与图均由随附代码重新计算得到；真实项目交付结构与此一致，内容按书面方案定制。